Exhibit 1

COMMISSIONER FOR PATENTS
UNITED STATES PATENT AND TRADEMARK OFFICE
WASHINGTON, D.C. 20231
www.uspto.gov

APPLICATION NUMBER	FILING DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	DRAWINGS	TOT CLAIMS	IND CLAIMS
60/246,052	11/06/2000		150	YOR920000785US1	7		

FILING RECEIPT

OC000000005685094

Robert P Tassinari
IBM Corporation Intellectual Property Law Dept
PO Box 218
Yorktown Heights, NY 10598

Date Mailed: 01/17/2001

Receipt is acknowledged of this provisional Patent Application. It will be considered in its order and you will be notified as to the results of the examination. Be sure to provide the U.S. APPLICATION NUMBER, FILING DATE, NAME OF APPLICANT, and TITLE OF INVENTION when inquiring about this application. Fees transmitted by check or draft are subject to collection. Please verify the accuracy of the data presented on this receipt. **If an error is noted on this Filing Receipt, please write to the Office of Initial Patent Examination's Customer Service Center. Please provide a copy of this Filing Receipt with the changes noted thereon. If you received a "Notice to File Missing Parts" for this application, please submit any corrections to this Filing Receipt with your reply to the Notice. When the PTO processes the reply to the Notice, the PTO will generate another Filing Receipt incorporating the requested corrections (if appropriate).**

Applicant(s)

Ana B Benitez-Jimenez, New York, NY ;
Chung-Sheng Li, Ossining, NY ;
John R Smith, New Hyde Park, NY ;

Continuing Data as Claimed by Applicant

Foreign Applications

If Required, Foreign Filing License Granted 01/17/2001

Title

Media net: a multimedia network for knowledge representation

Preliminary Class

Data entry by : STANBACK, PAUL **Team :** OIPE **Date:** 01/17/2001

LICENSE FOR FOREIGN FILING UNDER
Title 35, United States Code, Section 184
Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CRF 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 36 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Office of Export Administration, Department of Commerce (15 CFR 370.10 (j)); the Office of Foreign Assets Control, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

PLEASE NOTE the following information about the Filing Receipt:

- The articles such as "**a**," "**an**" and "**the**" are not included as the first words in the title of an application. They are considered to be unnecessary to the understanding of the title.
- The words "**new**," "**improved**," "**improvements in**" or "**relating to**" are not included as first words in the title of an application because a patent application, by nature, is a new idea or improvement.
- The title may be truncated if it consists of more than 600 characters (letters and spaces combined).
- The docket number allows a maximum of 25 characters.
- If your application was submitted under 37 CFR 1.10, your filing date should be the "date in" found on the Express Mail label. If there is a discrepancy, you should submit a request for a corrected Filing Receipt along with a copy of the Express Mail label showing the "date in."
- The title is recorded in sentence case.

Any corrections that may need to be done to your Filing Receipt should be directed to:

Assistant Commissioner for Patents
Office of Initial Patent Examination
Customer Service Center
Washington, DC 20231

Express Mail Label: EL480513501US
Date of Deposit: November 6, 2000

MediaNet: A Multimedia Network for Knowledge Representation

In this document, we present MediaNet, which is a knowledge representation framework that uses multimedia content for representing semantic and perceptual information, and the encoding, personalization, construction, and use of MediaNet in applications. The main components of MediaNet include conceptual entities, which correspond to world entities, and relationships among concepts. MediaNet allows the concepts and relationships to be defined or exemplified by multimedia content such as images, video, audio, graphics, and text. MediaNet models the traditional relationship types such as generalization and aggregation but adds additional functionality by modeling perceptual relationships based on feature similarity and constraints. For example, MediaNet allows a concept such as "car" to be defined as a type of a "transportation vehicle", but which is further defined and illustrated through example images, videos and sounds of cars. In designing the MediaNet framework, we have built on the basic principles of semiotics and semantic networks in addition to utilizing the audio-visual content description framework being developed as part of the MPEG-7 multimedia content description standard.

By integrating both conceptual and perceptual representations of knowledge, MediaNet has potential to impact a broad range of applications that deal with multimedia content at the semantic and perceptual levels such as querying, browsing, summarizing, and synthesizing multimedia. In particular, we have found that MediaNet can improve the performance of multimedia retrieval applications by using query expansion, refinement and translation across multiple content modalities. We report on experiments that use MediaNet in searching for images. We construct the MediaNet knowledge base using both WordNet and an image network built from multiple example images and extracted color and texture descriptors. Initial experimental results demonstrate improved retrieval effectiveness using MediaNet in a content-based retrieval system.

1 Introduction

Audio-visual content is typically formed from the projection of real world entities through an acquisition process involving cameras and other recording devices. In this regard, audio-visual content acquisition is comparable to the capturing of the real world by human senses. This provides a direct correspondence of human audio and visual perception with the audio-visual content [17]. On the other hand, text or words in a language can be thought of as symbols for the real world entities. The mapping from the content level to the symbolic level by computer is quite limited and far from reaching human performance. As a result, in order to deal effectively with audio-visual material, it is necessary to model real world concepts and their relationships at both the symbolic and perceptual levels.

In order to address the problem of representing real world concepts using semantics and perceptual features, we propose the MediaNet multimedia knowledge representation framework. MediaNet represents the world using concepts and relationships that are defined and exemplified using multiple media. MediaNet can be used to facilitate the extraction of knowledge from multimedia material and improve the performance of multimedia searching and filtering applications. In MediaNet, concepts represent world entities. Furthermore, relationships can be conceptual (e.g., Specializes) and perceptual (e.g., Is Similar To). The framework offers functionality similar to that of a dictionary or encyclopedia and a thesaurus by defining, describing and illustrating concepts, but also by denoting the similarity of concepts at the semantic and perceptual levels.

Previous work has focused on the development of multimedia knowledge bases for information retrieval such as the multimedia thesaurus (MMT) [21], a central component of the MAVIS 2 system [4], and visual pattern libraries [8]. The multimedia thesaurus provides concepts, which are abstract entities of the real world objects, semantic relationships such a generalization and specialization, and media representations of the concepts, which are portions of multimedia materials and associated features vectors. MediaNet extends the multimedia thesaurus this notion of relationships to include perceptual relationships that can also be exemplified and defined using audio-visual content. Furthermore, MMT treats semantic objects and perceptual information quite differently. MMT defines concepts that correspond to high-level, semantically meaningful objects in the real world with names in a language ("car" and "man") [6][22]. However, this exclude concepts that represent patterns based on perceptual information that are not named, such as the texture and color patterns in visual libraries [8].

By integrating both conceptual and perceptual representations of knowledge, MediaNet has potential to impact a broad range of applications that deal with multimedia content at the semantic and feature levels such as query, browsing, summarization, and synthesis. In particular, we have found that MediaNet can improve the performance of multimedia retrieval applications by using query expansion, refinement and translation across multiple content modalities. In this paper, we report on experiments that use MediaNet in searching for images. We construct the MediaNet knowledge base using both WordNet and an image network built from multiple example images and extracted color and texture descriptors. Initial experimental results demonstrate improved retrieval effectiveness using MediaNet in a content-based retrieval system; however, more extensive experiments are needed.

This document is organized as follows. In section 2, we describe the main components of MediaNet. Section 3 describes the encoding of MediaNet using MPEG-7 description tools. The personalization, construction, and use of MediaNet in applications are discussed in sections 4, 5, and 6, respectively. Section 7 describes the implementation of an extended content-based retrieval system, which uses the MediaNet framework and reports on the experiments that compare the performance of the extended content-based retrieval system to a typical content-based retrieval system. Finally, section 8 concludes summarizing the results of this work.

2 MediaNet

MediaNet is a multimedia network information framework for representing world knowledge through both symbolic and multimedia information (e.g., images and video sequences), including feature descriptors. In MediaNet, world entities are represented using concepts and relationships among concepts, which are defined and exemplified by multimedia information and relationships among multimedia information. For example, the concept Human can be represented by the words "human", "man", and "homo", the image of a human, and the sound recording of a human talking, and can be related to the concept Hominid by a Specialization and Similar Shape relationships; the concept

Hominid can be represented by the word "hominid" and the text definition "a primate of the family Hominidae" (see Figure 1). In this section, we describe the main components of MediaNet.

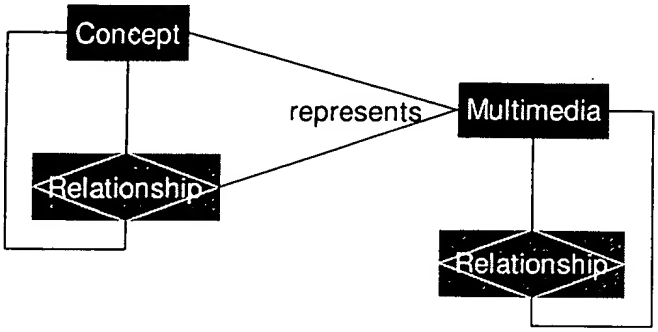


Figure 1: Relations among the main components of MediaNet.

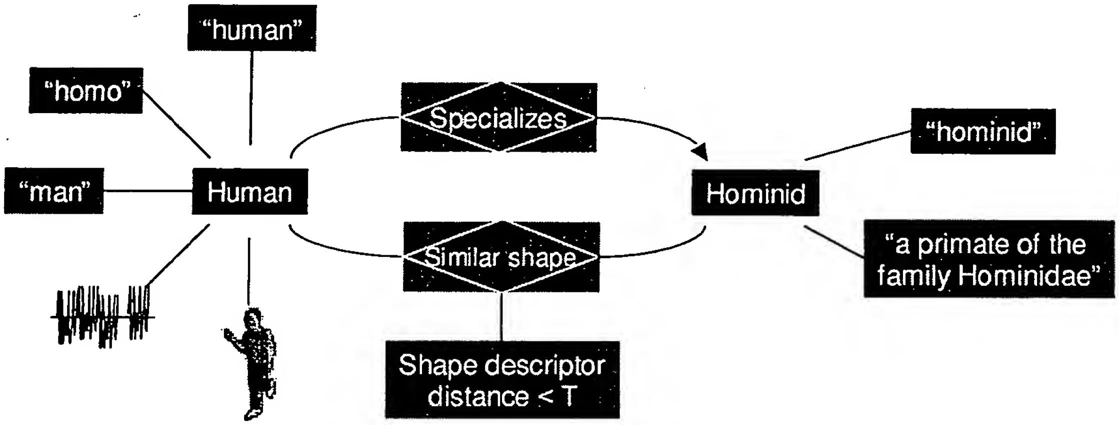

"human"
"homo"
"man"
Human
Specializes
Similar shape
Shape descriptor distance < T
Hominid
"hominid"
"a primate of the family Hominidae"

Figure 2: Concept Human and Hominid

2.1 World

MediaNet can represent the knowledge of the real world in which we humans live or the fictional worlds staged in movies. The knowledge of each world could be described in a separate MediaNet knowledge base. The knowledge of multiple worlds could be integrated in one MediaNet knowledge base if different contexts are defined in MediaNet as in [7].

2.2 Concepts

Concepts are the basic units for knowledge representation in MediaNet. Concepts represent abstractions of entities in the world such as inanimate objects, living entities, events, properties, and sensations evoked by the previous. Examples of concepts are Human (see Figure 1) and Car that refer to any living entity human and any inanimate object car, respectively; Wedding is the concept of an event; and Blue can be a property or sensation concept. Concepts can refer to classes of entities in the world such as Car; unique and identified entities such as Ronald Reagan; and abstract entities with no physical presence in the world such as Freedom.

Concepts can be categorized into semantic and perceptual. Semantic concepts correspond to semantically meaningful entities in the world with names in a language such as Car. On the other hand, perceptual concepts can be fully defined using impressions obtained by the use of senses and not be named in a language such a specific texture pattern. In particular, low-level audio-visual feature descriptors such as color histogram and Tamura texture can be used to define audio-visual concepts.

There are important differences between words in English and other languages and concepts. In the previous examples, concepts were usually named with one word; however, there may be no words to designate some concepts, more than one word to designate the same concept, or no words to uniquely designate a concept. An example of the first case is the texture of a specific piece of fabric. The second case corresponds to synonyms, i.e., words having the same or nearly the same meaning or sense in a language, as "human" and "man" (see Human concept in Figure 1). Finally, the third case corresponds to polysemy, i.e., a word having multiple meanings in a language, as "studio" which can be a "the workroom or atelier of an artist", "a room or set of rooms specially equipped for broadcasting radio or television programs, etc.", and "an apartment consisting of one main room, a kitchen, and a bathroom".

2.3 Concept Relationships

Concepts can be related to other concepts by semantic and perceptual relationships. Semantic relationships relate concepts based on their meaning (e.g., specializes and opposite to); perceptual relations relate concepts based on perceptions concept entities (e.g., similar shape and darker). In Figure 1, *Specializes* and *Similar Shape* are semantic and perceptual relations between Human and Hominid, respectively.

The relations in traditional thesaurus such as the Dewey Decimal System [3] and lexical systems such as WordNet [10] are semantic relations. The most common relationships among terms in traditional thesauri are generalization/specialization, equivalence, and related. All the semantic relationships in WordNet except for synonymy apply between concepts in MediaNet; these are listed in Table 1 together with definitions and examples. Synonyms are text representations of the same concept in MediaNet (see Figure 1) so it can not relate concepts. Antonymy is a binary, symmetric relationship between concepts; the rest of the relationships in Table 1 are binary and transitive. There is usually one specialized concept (hypernymy) per concept so this relationship organizes the concepts into a hierarchical structure.

Table 1: Definitions and examples of the semantic relationships in WordNet except for synonymy.

Relationship	Definition	Example
Antonymy	To be opposite to	White *Is Opposite To* Black
Hypernymy/ Hyponymy	To specialize To generalize	Hominid *Specializes* Human Human *Generalizes* Hominid
Meronymy/ Holonymy	To be a part, member, or substance of To have a part, member, or substance of	Ship *Is Member Of* Fleet Martini *Has Substance* Gin
Entailment	To entail or, to cause or involve by necessity or as a consequence	Divorce *Entails* Marry
Troponymy	To be a manner of	Whisper *Is Manner Of* Speak

Concepts refer to entities that are perceived by senses or evoked from these perceptions. Therefore, concepts can also be related by perceptual relationships. Examples of perceptual relationships are visual relationships (e.g., Sky *Has Similar Color To* Sea and Human *Has Similar Shape To* Hominid; see Figure 1) and audio relationships (e.g., Stork *Has Similar Sound To* Pigeon; see Figure 3). Audio-visual relationships are special because they are recorded in the audio-visual content and, therefore, can have multimedia representations as concepts.

2.4 Multimedia Representations

Concepts and concept relationships may be illustrated by multimedia information such as images, video program, and feature descriptors, among others. As an example, the concept Human is exemplified by the words "human", "man", and "homo", the image of a human, and the sound recording of a human in Figure 1.

Multimedia representations are not necessarily whole multimedia documents such as an image; they can be sections of multimedia material and have associated feature descriptors and relationships extracted from the multimedia content. A media representation of the concept Human can be a region of an image that depicts a human and the value of a contour shape descriptor for that region. A concept can also be illustrated by the value of a feature descriptor with no multimedia material. As an example, the concept Human can be presented by the value of a contour shape descriptor. The value of the contour shape feature may be the average of the contour shape feature values for a set of image regions depicting a human; however, the image data may not be included in the media representation only the value of the contour shape descriptor. The same applies to any other media such as text, audio, and video.

Some types of multimedia information may not be relevant or applicable to represent a concept. Cars can be of many colors; therefore, color histogram will not be useful to represent the concept Car. Audio is not relevant to represent the concept Blue; visual descriptors do not apply to the concept Jazz either. Section 4 describes how weights can be assigned to each representation to indicate how relevant the representation is for a concept.

Audio-visual relationships between concepts are recorded in the audio-visual content and, therefore, can have multimedia representations as concepts. These relationships can usually be generated and inferred by automatic tools and expressed as constraints on and similarity of audio-visual feature descriptors. They can also be exemplified by audio-visual content related with that relationship, as an example, Figure 3 exemplifies the relationship *Has Similar Sound* using two image regions.

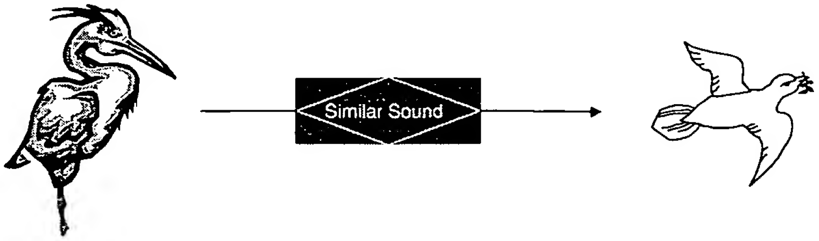


Figure 3: Example of audio relationship *Has Similar Sound.*

We shall provide some more examples of media representations of concepts now. The concept Car can have the following media representations: the word "car", the text definition "an automobile", an image depicting a car together with shape features extracted from it, and the sound recording of a running car. The concept Blue can have the following representations: the English word "blue", the Spanish word "azul", the text definition "the pure color of a clear sky; the primary color between green and violet in the visible spectrum", and the value of the color histogram corresponding to blue. Text representations may be in different languages.

2.5 Multimedia Relationships

Multimedia information such as audio-visual content and feature descriptors can be related by audio-visual relationships as concepts. An example of visual relation between two image regions is shown in Figure 3.

3 Encoding of MediaNet Using MPEG-7

The MPEG-7 standard [13] aims at standardizing tools for describing the content of multimedia material in order to facilitate a large number of multimedia searching and filtering applications. In this section, we describe and give examples on how MediaNet knowledge bases could be encoded using MPEG-7 description tools, which would greatly benefit the exchange and re-use of multimedia knowledge among applications.

Relevant MPEG-7 description tools for encoding MediaNet knowledge bases are the CollectionStructure DS and the Semantic DS [11][12]. The CollectionStructure DS allows describing collections of multimedia content, probability information for collections, and relations between collections. The Semantic DS allows representing the entities existing in the world depicted by audio-visual content and the relationships among the world entities. Below, we describe the mapping of MediaNet's components to the CollectionStructure DS and the Semantic DS. The preferred encoding of MediaNet is using the Semantic DS because it is the more direct and intuitive mapping.

It is important to note that the MPEG-7 description tools are still under development within the MPEG-7 standard and their specification may change in the near future and, therefore, the encoding MediaNet knowledge bases with MPEG-7 description tools.

3.1 CollectionStructure DS

The CollectionStructure DS allows describing collections of content, probability distribution for collections based on feature descriptors, (e.g., feature descriptor centroids) and relationships (e.g., Specializes) between collections. The elements of a collection can be images, videos, and portions of audio-visual content, among others. A collection can be described by text annotations, among others.

The main components of MediaNet could be mapped to the CollectionStructure DS as follows: a MediaNet knowledge base to a collection structure, each concept to a content collection, and each concept relationship to a collection relationship. The text representations of a concept could be encoded as text annotations of the collection; other multimedia representations (e.g., images and features) could be described as the elements or probability distributions of the collection. The XML description of the example in Figure 1 is included below. We assume the reader is familiar with the markup language XML [23].

```
<CollectionStructure id="MediaNet0">
        <Collection xsi:type = "mpeg7:ContentCollectionType" id = "ConceptHuman"> <!--- Concept Human -->
                <TextAnnotation><FreeText> human </FreeText></TextAnnotation>
                <TextAnnotation><FreeText> homo </FreeText></TextAnnotation>
                <TextAnnotation><FreeText> man </FreeText></TextAnnotation>
                <Image>
                        <MediaLocator> <MediaURI>Human.jpg</MediaURI> </MediaLocator>
                </Image>
                <Audio>
                        <MediaLocator> <MediaURI>Human.wav</MediaURI> </MediaLocator>
                </Audio>
        </Collection>
        <Collection xsi:type = "mpeg7:ContentCollectionType" id = "ConceptHominid"> <!--- Concept Hominid -->
                <TextAnnotation><FreeText> hominid </FreeText></TextAnnotation>
                <TextAnnotation><FreeText> a primate of the family Hominidae </FreeText></TextAnnotation>
        </Collection>
        <!-- Graph describing Specializes relationship from concept Human to concept Hominid -->
        <Graph>
                <Edge name="Specializes" source="ConceptHuman" target="ConceptHominid"/>
        </Graph>
</CollectionStructure>
```

3.2 Semantic DS

The Semantic DS allows describing the semantic entities in the world and relations among the semantic entities. The semantic entities include textual labels and have associated analytical models that can be descriptor collection models, content collection models, and probability distribution models.

The main components of MediaNet could be mapped to the Semantic DS as follows: a MediaNet knowledge base to a semantic description, each concept to a semantic entity, and the concept relationships to relations among semantic entities. The text representations of a concept could be described as text labels of the semantic entity; other multimedia representations of a concept (i.e., multimedia content, feature descriptors, and descriptor statistics) could be described as analytical models of the semantic entity (i.e., content collection models, descriptor collection models, and probability distribution models, respectively). The XML description of the example in Figure 1 is included below. We have considered the concepts Human and Hominid also are represented by the centroid of a contour shape descriptor.

```
<Semantic id="MediaNet0">
        <SemanticBase id= "ConceptHuman"> <!--- Concept Human -->
                <Label><FreeTerm> human </FreeTerm></Label>
                <Label><FreeTerm> homo </FreeTerm> </Label>
                <Label><FreeTerm> man </FreeTerm> </Label>
                <AnalyticalModel xsi:type = "mpeg7:ContentCollectionModel">
                        <Collection xsi:type = "mpeg7:ContentCollectionType">
                                <Image>
                                        <MediaLocator> <MediaURI>Human.jpg</MediaURI> </MediaLocator>
                                </Image>
                                <Audio>
                                        <MediaLocator> <MediaURI>Human.wav</MediaURI> </MediaLocator>
                                </Audio>
                        </Collection>
                </AnalyticalModel>
                <AnalyticalModel xsi:type = "mpeg7:ProbabilityModelClass"> <!--  Centroid of countour shape descriptor  -->
                        <DescriptorName> ContourShape </DescriptorName>
                        <ProbabilityModel xsi:type = "ProbabilityDistribution" dimensions = "64">
                                <Mean dim = "64">20 6 26 17 53.5 1 41 4.5 20.5 0 2 1 2 1 2 2 23.5 1.5 26 22 69 0 116.5
40.5 25 0 3 2.5 1 0 0 1 34.5 2 11 25.5 74 0 4.5 34 6 0 0 2.5 0 0 0 7 33 36.5 147.5 30 27 0 0 11.5 0 0 0 10.5 0 0 0 11.5 </Mean>
                        </ProbabilityModel>
                </AnalyticalModel>
        </SemanticBase>
        <SemanticBase id = "ConceptHominid"> <!--- Concept Hominid -->
                <Label><FreeTerm> hominid </FreeTerm></Label>
                <Label><FreeTerm>a primate of the family Hominidae</FreeTerm></Label>
                <AnalyticalModel xsi:type = "mpeg7:ProbabilityModelClass"> <!--  Centroid of countour shape descriptor  -->
                        <DescriptorName> ContourShape </DescriptorName>
                        <ProbabilityModel xsi:type = "ProbabilityDistribution" dimensions = "64">
                                <Mean dim = "64">20 6 26 11 53.5 1 41 4.5 20.5 0 2 1 2 1 2 2 23.5 1.5 26 22 69 0 116.5
40.5 25 0 3 2.5 1 0 0 1 34.5 2 11 25.5 74 0 4.5 34 6 0 0 2.5 0 0 0 0 33 34.5 147.5 30 27 0 0 11.5 0 0 0 10.5 0 0 0 11.5 </Mean>
                        </ProbabilityModel>
                </AnalyticalModel>
        </SemanticBase>
        <!-- Graph describing Specializes relationship from concept Human to concept Hominid -->
        <Graph>
                <Edge name="Specializes" source="ConceptHuman" target="ConceptHominid"/>
        </Graph>
</Semantic>
```

4 Personalization of MediaNet

A MediaNet KB is composed of content, relationship and multimedia nodes, and arcs among them. The personalization of a MediaNet knowledge base consists on assigning weights to arcs or nodes of a MediaNet knowledge base. Weights can be generated during the construction or the usage of a MediaNet knowledge base.

4.1 Construction

Automatic, semi-automatic, and manual methods can be used to construct a MediaNet knowledge base, all of which are subject to error. Therefore, it is important in using MediaNet to have an indication of how correct and representative of the underlying world the knowledge base is. For example, the construction weight of an arc between a concept and an image may indicate the probability that the image is indeed an example of the concept or how well the image represents the concept.

4.2 Usage

Applications using MediaNet can also keep and update weights of arcs and nodes to personalize the use of the MediaNet for a specific task such as query or browsing. The weights can be learned and updated from users through relevant feedback. The feedback provided by users would vary from application to application and may be direct or indirect. Direct feedback involves the user explicitly specifying the relevance of the results returned by an application task, for example, the user could label the results of a query as being positively or negatively relevant to the query. Indirect feedback can be deduced by the actions performed by users in response to the results of an application task. For example, if a web search engine returns links to four web pages as a result of a search, the web pages visited by the user could be considered closer or more relevant to the query and the weights could be updated accordingly.

5 Construction of MediaNet

In constructing a MediaNet knowledge base, we distinguish between creating the knowledge base from scratch and updating an existing knowledge base by adding new concepts, illustrative multimedia information, and/or relationships.

5.1 Creation

A MediaNet knowledge base for a specific domain can be created starting with collection of representative multimedia information such as images, video programs, audio sequences, textual annotations, and/or features descriptors. Existing thesaurus and knowledge bases, automatic extraction tools, automatic visual classification systems, and human assistance can be combined to create a MediaNet knowledge base.

The creation of a MediaNet knowledge base can be decomposed into the following steps:

- Create the multimedia network(s)

The extraction of feature descriptions from multimedia material can be done using automatic feature extraction tools. Relationships of feature similarity among multimedia information (e.g., Sound Similar) can be automatically computed from feature descriptor values using distance functions such as Euclidean and L1 distances. Other relationships among multimedia information can be expressed by constraints on feature descriptors (e.g., A Left Of B as A.x < B.x). A library of relevant constraints for the application could be created for a human expert or be learned.

- Create the perceptual concept network(s)

Perceptual concepts can be created by clustering or quantizing multimedia information, among others. If we organize a collection of multimedia material into clusters based on their associated feature descriptors, the resulting clusters could be represented as perceptual concepts in MediaNet as done in [8] to create a thesaurus of texture patterns. Another method of generating perceptual concepts is to quantize the feature space of a feature descriptor to create a library of color and texture patterns.

- Create the semantic concept network(s)

The network of semantic concepts and semantic relationships among the concepts with corresponding text representations could be designed manually by human experts in the domain/world being represented. This is usually the only way to create a knowledge base for very specialized domains such as medical and research images. However, there are an increasing number of thesauri and knowledge bases available in electronic form such as WordNet and the Dewey Decimal System that can be used to automatically or, at least, semi-automatically create semantic network(s) for more general domains. Some of these knowledge bases were designed for specific domains such as art images. Specialized knowledge bases can be combined with more general knowledge bases such as WordNet to satisfy the requirements of many applications. Applications are not likely to handle all the concepts in one or more existing knowledge bases but to select and use portions of existing knowledge bases that are most relevant for the domain.

- Link the multimedia, the perceptual concept, and the semantic concept network(s) to form one network

This process is likely to be done or supervised by humans because it is one of the most sensitive steps in creating a MediaNet knowledge base. After a training phase, automatic classification tools based on feature descriptors could be used to link multimedia information, perceptual concepts, and semantic concepts. If text annotations are available for multimedia material, these can be used to train classifiers automatically.

- Create other nodes in the network

When multimedia information, perceptual concepts, and semantic concepts are linked together additional nodes can be added to the MediaNet knowledge base such as statistics of feature descriptors (e.g., centroid) for the multimedia examples of a concept.

5.2 Update

This process consists on adding new multimedia information, concepts, and/or relationships to a MediaNet knowledge based. New multimedia information could be linked to concepts in a similar way to step fourth of the creation of a MediaNet knowledge base. Changes in the topology of the network of MediaNet can be the response to feedback from users and be done automatically using techniques such as self-organizing maps especially for perceptual concepts. Human experts can also manually edit and add new concepts and/or relationships among concepts to a MediaNet knowledge based as needed.

6 Usage of MediaNet

This section describes how to perform query, browsing, summarization, and synthesis of multimedia using a MediaNet knowledge base with learned usage weights. For all the applications, network of content, relations, and multimedia information in MediaNet is mapped into one or more metric spaces to evaluate the distance between two nodes in the network.

6.1 Query

The objective of querying a MediaNet knowledge base is to find the most relevant grouping of concepts and/or multimedia information to an input query by the user. The interest of users may change over time. We can assume that the query is specified using the MediaNet's constructs: concepts, relationships, and multimedia information. Methods for calculating distance measures in a metric space, finding the minimum cost path between two nodes in a network, and graph matching are used in this application.

The query of a MediaNet knowledge base can be decomposed into the steps described below. The MediaNet knowledge base and the query are viewed as a set of nodes and arcs between the nodes. Nodes can be perceptual concepts, semantic concepts, and multimedia information nodes (multimedia content or feature descriptors). A query weight between two connected nodes in the network indicates how relevant is to consider the connection between the two nodes when calculating distances to other nodes.

- Initialize query weights for arcs in the NediaNet network

The initial query weights can be set initially to the same value or random values. These weights can then be successively updated through interaction with the user to reflect the user's changing interest.

- Calculate the direct distance between one pair of nodes in a network

The feature distance between two feature descriptor nodes can be calculated using any standard distance function. Feature descriptors should be normalized so that distances fall within the interval [0,1]. The feature distance between two multimedia content nodes can be calculated as the weighted sum of distances for the common feature descriptors. The query weights of the feature descriptors for each multimedia content nodes can be considered. The

direct distance between two concept nodes is the sum of their feature and semantic distances. The feature distance between two concept nodes can be calculated as the weighted sum of distances for common feature descriptors of the concepts and associated multimedia content nodes. The query weights of the feature descriptors for the multimedia content nodes and the concept nodes can be considered.
Concept nodes can also be related by semantic relationships. The semantic distance between two concept nodes can be calculated as the weighted sum of distances of direct semantic relations between the two conceptual nodes. The query weights of the semantic relationships are considered. However, multimedia content nodes can also be related to other multimedia content nodes and concept nodes, which can be considered while calculating the distance between two multimedia content nodes. As feature descriptor, multimedia content, and concept nodes can be related among each other with a network, potentially, two arbitrary nodes in the network could be compared by matching the networks centered at each node, which is a very expensive process. Only specific cases such as the semantic relations among concepts could be taken into account to simplify the problem.

- Find minimum distance between each pair of nodes in the MediaNet network

Not all the possible connections between nodes in the MediaNet network are explicit in the network. An implicit arc between two nodes of the same type is considered if a distance measure can be obtained for them. Two nodes in the MediaNet network may be connected by more than one path in the network. This step involves finding the minimum distance between each pair of nodes through all the alternative paths between them. There are very well known methods to find the minimum cost (distance) and the corresponding path between two nodes in a graph if the cost (distance) between each pair of two nodes is already computed (see previous step).

- For each node in the query network, find minimum distance to each node of same type in the MediaNet network

This step follows the same procedure as step number one. Only nodes of the same type (e.g., only feature descriptor nodes are compared with feature descriptor nodes) or nodes of any type may be matched in this process (e.g., feature descriptors and multimedia content nodes are matched).

- Join the results for each node in the query network and obtain global distance measure

Join the results for each node in the query network to satisfy as close as possible the relationships among the nodes in the query network. Graph matching algorithm can be used in this process.

- Update relative and global browsing weights based on feedback from users

Direct and indirect feedback can be continuously collected from the users to update the query weights. An example of direct feedback from a user would be the user indicating the degree of relevancy of each node returned by the query. An example of indirect feedback from a user would be the user continuing the search with a subset of the nodes returned by the query.

6.2 Browsing

The objective of browsing the MediaNet knowledge base is to navigate the most relevant concepts and/or media representations to a user. The interest of users may change over time. Methods for visualizing network of nodes and relevance feedback are used in this application.

The browsing of a MediaNet knowledge base can be decomposed into the steps described below. The MediaNet knowledge base is viewed as a set nodes and arcs between the nodes. Nodes can be perceptual concepts, semantic concepts, and multimedia information nodes (multimedia content or feature descriptors). Two types of browsing weights can be maintained: global weights assigned to individual nodes or group of nodes in the network that indicate the global importance of those nodes in the browsing; and relative browsing weights between two nodes connected by the network that indicate the relative importance of displaying one node if the other node is displayed, i.e., how important is for both nodes to be displayed at the same time.

- Initialize the global browsing weights for browsing in the NediaNet network

This process may involve the user to query the MediaNet knowledge base as described in the previous section. The top results can be considered the areas (group of network nodes) of interest for browsing. The global browsing weights for these areas can be initialized with the distance scores returned by the query. If no initial query is performed, then, the same value or random values can be assigned to the global weights of nodes in the MediaNet network. These weights can then be successively updated through interaction with the user to reflect the user's changing interest.

- Initialize the relative browsing weights of nodes for browsing in the NediaNet network

The relative browsing weight two connected nodes can be set initially to the minimum distance between the two nodes. These weights can then be successively updated through interaction with the user to reflect the interest of users.

- Select relevant nodes surrounding area(s) of interest in the MediaNet network

A MediaNet network can consist of many nodes. Global and relative browsing weights are used to determine what nodes are going to be displayed and how during browsing. Nodes in the network with higher global weights are displayed before areas with lower global weights. Nodes with higher relative weights for the nodes in the areas of interest are displayed before nodes with lower relative weights.

- Display relevant nodes in the MediaNet network

The decision on how and where the nodes are going to be displayed to the user on the screen can be based on the global, relative, and minimum distances between the nodes. Several visualization techniques could be used to achieve this.

- Update relative and global browsing weights based on feedback from users

Direct and indirect feedback can be continuously collected from the users to update the global and relative browsing weights. An example of direct feedback from a user would be the user selecting sets of displayed nodes and indicating his degree of interest for each of them. An example of indirect feedback from a user would be the user requesting more or less detail about a group of displayed nodes.

6.3 Summarization

The objective of summarizing multimedia material using a MediaNet knowledge base is to select the most relevant section of multimedia content to the user. The interest of users may change over time. Methods for segmentation and extraction of key representations for multimedia material, automatic classification, and relevance feedback are used are used in this application.

The summarization of multimedia using a MediaNet knowledge base can be decomposed into the steps described below. The MediaNet knowledge base is viewed as a set nodes and arcs between the nodes. Nodes can be perceptual concepts, semantic concepts, and multimedia information nodes (multimedia content or feature descriptors). Two types of summarization weights can be maintained: global weights assigned to individual nodes or group of nodes in the network that indicate the global importance of those nodes to generate the summary; and relative weights between two nodes connected by the network that indicate the relative importance of one node relevant to generate the summary if the other node is selected, i.e., how important is for both nodes to participate in generating the summary at the same time.

- Initialize the global and relative summarization weights in the NediaNet network

The same procedure described for the first two steps of browsing applies here.

- Select relevant criteria to segment and select key representations for multimedia material

Feature descriptors can be extracted from the multimedia material to be summarized. These features can be used to initially classify the multimedia material into one or more concepts in the MediaNet knowledge base. The most representative feature descriptors and descriptor values for those concepts can be selected as the criteria to decide how to segment the multimedia material (e.g., segment a video program into shots using color and motion) and to select appropriate key representations of each section of the segmented multimedia material (e.g., select the key frame in a video shot as the frame whose color is not similar to the representative one).

- Segment and select key representations for multimedia material

Existing methods can be used to segment video, audio, image, and multimedia documents into homogeneous sections of content and to select key representations for each section. As described in the previous step, this could be done using the MediaNet knowledge base too. Each section of content resulting from the segmentation process could also classified into one or more concepts in the MediaNet knowledge based to select more specific criteria to select appropriate key representations of the sections (e.g., the regions resulting from segmenting an image are classified into concepts to determine the most relevant key feature to represent them).

- Construct the summary based on key representations

The decision on how and when/where the key representations are going to be used to generate the final summary can be based on the global, relative, and minimum distances between concepts corresponding to the key representations. Some of the parameters would need to be set in the summary are the summary length, the display order of key representations, etc.

- Update relative and global synthesis weights based on feedback from users

The same procedure followed for the last step of browsing applies here.

6.4 Synthesis

The objective of the synthesis of multimedia material using a MediaNet knowledge base is to create a multimedia overview of a portion of the MediaNet knowledge base. The interest of users may change over time are used in this application.

The synthesis of multimedia from a MediaNet knowledge base can be decomposed into the steps described below. The MediaNet knowledge base is viewed as a set nodes and arcs between the nodes. Nodes can be perceptual concepts, semantic concepts, and multimedia information nodes (multimedia content or feature descriptors). Two types of synthesis weights can be maintained: global weights assigned to individual nodes or group of nodes in the network that indicate the global importance of those nodes in the synthesis; and relative weights between two nodes connected by the network that indicate the relative importance of synthesizing one node if the other node is synthesized, i.e., how important is for both nodes to be synthesized at the same time.

- Initialize the global and relative synthesis weights in the NediaNet network

The same procedure described for the first two steps of browsing applies here.

- Select relevant nodes within or surrounding the portion to the knowledge base to be synthesized

A MediaNet network can consist of many nodes. Global and relative browsing weights are used to determine what nodes within a portion of the knowledge base are going to be synthesized and how. Nodes in the network with higher global weights are synthesized before areas with lower global weights. Nodes with higher relative weights for the nodes in the areas of interest are synthesized before nodes with lower relative weights.

- Synthesize relevant nodes in the MediaNet network

The decision on how and where the nodes are going to be displayed to the user on the screen can be based on the global, relative, and minimum distances for nodes. Some of the parameters would need to be set in the synthesis of network nodes are the displaying time, the size of the display area over time, the most representative multimedia content for each node, etc.

- Update relative and global browsing weights based on feedback from users

The same procedure followed for the last step of browsing applies here.

7 Implementation and Experiments of MediaNet in CBIR System

MediaNet extends current knowledge representation frameworks by including multimedia information. By integrating both conceptual and perceptual representations of knowledge, MediaNet has potential to impact a broad range of applications that deal with multimedia content at the semantic and feature levels. In this section, we describe the implementation of an extended content-based retrieval system that uses MediaNet to expand, refine, and translate queries across multiple content modalities. In this section, we focus on describing the construction and use of the MediaNet knowledge base in a content-based retrieval system.

Typical content-based retrieval systems index and search image and video content by using low-level visual features (for example, see [1][5][18]). These systems automatically extract low-level features from the visual data, index the extracted descriptors for fast access in a database, and query and match descriptors for the retrieval of the visual data. The extended content-based retrieval system is a typical content-based retrieval system that also includes a MediaNet knowledge base with media representations of concepts and a query processor that uses the MediaNet knowledge base to refine, extend, or translate user queries (see Figure 4).

In the extended content-based retrieval system, we use color and texture features. The color descriptors are color histogram and color coherence; the texture descriptors are wavelet texture and tamura texture. Queries to the CB search engine can only be visual queries as the name of an image in the database or the value for any of the low-level features in the image database. The CB search engine uses weighted Euclidean distance function to obtain distance scores between the query and each image in the database to the query and returns an ordered list of images based on the distance scores. No text annotations are stored in the database or are used in the retrieval.

7.1 Construction

A MediaNet knowledge base could be created manually or automatically using classification, machine learning, and artificial intelligence tools among others. The MediaNet knowledge base for the extended content-based retrieval system was created semi-automatically using existing text annotations for some of the images in the database and the electronic lexical system WordNet. In this section we describe the procedure followed to construct the MediaNet knowledge base. More details of this procedure are included in Annex A.

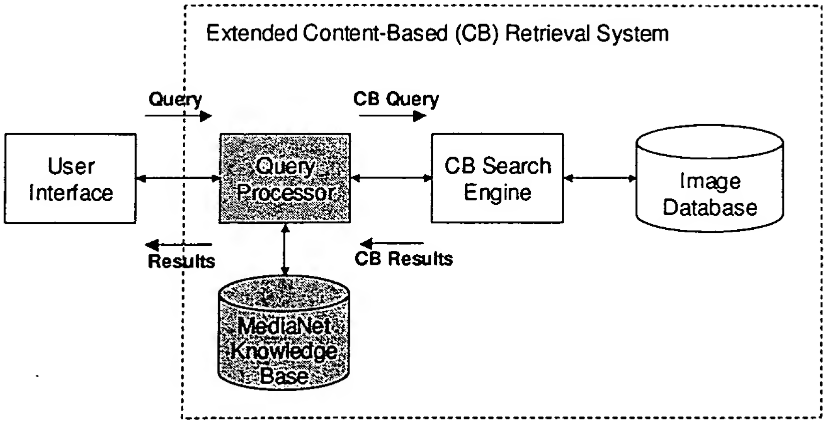


Figure 4: Components of the extended content-based retrieval engine. The new components with respect to typical content-based retrieval engines are shown in gray.

WordNet [10] is an electronic lexical system that organizes English words into set of synonyms, each representing a lexicalized concept. Semantic relationships link the synonym sets. The semantic relationships between words and words senses incorporated by WordNet include synonymy, antonymy, hypernymy/hyponymy, meronymy/holonymy, entailment, and troponymy (these relationships are defined in Table 1 except for synonymy). WordNet contains more than 118,000 different word forms and more than 90,000 word senses. Being available in electronic form, it is one of the most convenient tools for generating concepts, text representations of concepts, and relationships among concepts at the semantic level.

The first step for constructing the MediaNet knowledge was to create concepts and text representations of the concepts using WordNet and human assistance. First, the text annotations were serialized into words that were ordered alphabetically. Dummy words such a prepositions and articles and duplicated words were removed. Then, WordNet was used to generate the senses and the synonyms for each word. A human supervisor selected the correct sense (or senses) of each word in the context of the text annotations and the image data. As an example, for the annotation "singer in studio", the correct sense for the word "studio" was selected as "a room or set of rooms specially equipped for broadcasting radio or television programs, etc". The supervisor also specified the syntactic category for each word (noun, verb, etc.). A concept was created for each word/sense pair, and was assigned the sense and synonyms provided by WordNet as text representations.

The next step was to generate relationships among concepts. We decided to use the top three relationships listed in Table 1. We used WordNet to automatically generate all of the antonyms, the hypernyms/hyponyms, and the meronyms/holonyms for each concept (i.e., each word-sense pair), automatically parsed the output of WordNet, obtained the relationships among the concepts, and stored them in the MediaNet knowledge base.

Finally, visual representations of concepts were generated automatically using color and texture feature extraction tools. For all the images associated with a concept, we extracted color and texture features and computed the feature centroids (centroid for group of images). The visual representation of each concept was the list of images for the concept with associated feature values, and the feature centroids.

For each application, the list of concepts and relationships in the MediaNet knowledge base should be representative of the content in the database and the goal of the application task. We used the textual annotations already available for the images, which were quite brief and general. More specific and representative text annotations could have been produced and used to construct a more optimized knowledge base. The process of generating the concepts from the words in the textual annotations could be automated by processing the text annotations using natural language techniques, or by using latent semantic analysis to match each word and surrounding words in the annotations to the different senses of the word provided by WordNet, among others.

More advanced techniques could have been used to generate more suited visual representations of the concepts. Some ideas are selecting more than one feature representation for each concept using Kohonen feature map on the feature vectors of the images associated to the concept [8], latent semantic analysis techniques applied to feature vectors as in [24], assigning weights to the representations of concepts, and segmenting and extracting features of images at the region level instead of at the image level [14].

7.2 Usage

The extended content-based retrieval system is a typical content-based retrieval system that also includes a MediaNet knowledge base with media representations of concepts and a query processor that uses the MediaNet knowledge base to refine, extend, and translate user queries. Although the CB search engine only provides content-based retrieval based on color and texture features, the retrieval engine can accept text queries because MediaNet can be used to translate them into visual data. This is an added functionality to the retrieval engine provide by MediaNet. In this section, we shall describe how MediaNet is used by the query processor module to process queries in the different media integrated into the MediaNet knowledge base, in our case, visual and text queries. More details of this procedure are included in Annex B.

When the user submits a query, either textual or visual, the query processor identifies relevant concepts to the query together with importance weights. This is done by matching the query to the media representations of the concepts in the knowledge base (e.g., a text query is matched to text representations of the concepts) and obtaining a relevance score for each concept indicating the similarity of the media representations of the concept to the query. The top ranked concepts and semantically similar concepts to these are considered relevant for the query. Then, the query and the media representations of the relevant concepts are matched to the image database by sending multiple queries to the CB search engine. Finally, the results are collected and merged into a unique list to display them to the user. The procedure is shown in Figure 5.

For image queries, color and texture feature vector are extracted from the query image and matched against the feature centroid of each concept using Euclidean distance. Only the percentage P% of the top ranked concepts are considered relevant to the query and selected. The feature vectors for all the images associated to each relevant concept are then matched to the query. Again, the percentage Q% of the top ranked images are considered relevant to the query and selected. The final dissimilarity score of each concept to the query is calculated as a function of the distances of the feature centroid of the concept to the query, the average distance of the relevant images of the concept to the query, and the proportion of the images of the concept that are relevant by the following equation:

$$fdist(q,c) = dist(q,cen_c) + \sum_{i \in rel_c} \frac{dist(q,i)}{num\_rel_c} * \sqrt{\frac{num_c}{num\_rel_c}} \qquad (1)$$

where q is the feature vector of the query image, c is a relevant concept, cen_c is the feature centroid of concept c, num_c is the number of images of concept c, num_rel_c is the number of relevant images for concept c, rel_c are the feature vectors of the relevant images of concept c, i is a feature vector of one relevant image of concept c, and dist(q,c) is the weighted Euclidean distance among feature vectors q and c.

The top N most similar concepts to the query are then selected. The set of relevant concepts is expanded with at most M more concepts from the MediaNet knowledge base with lowest average conceptual dissimilarity to the relevant concepts. Only concepts with dissimilarities below R% the average conceptual distance in MediaNet can be selected as relevant. We only expand the set of relevant concepts with concepts that are similar enough semantically. For each concept added in this step, distance scores to the query are calculated as the sum of the average conceptual dissimilarity to the initial set of relevant concepts and the average distance of the initial set of relevant concepts to the query.

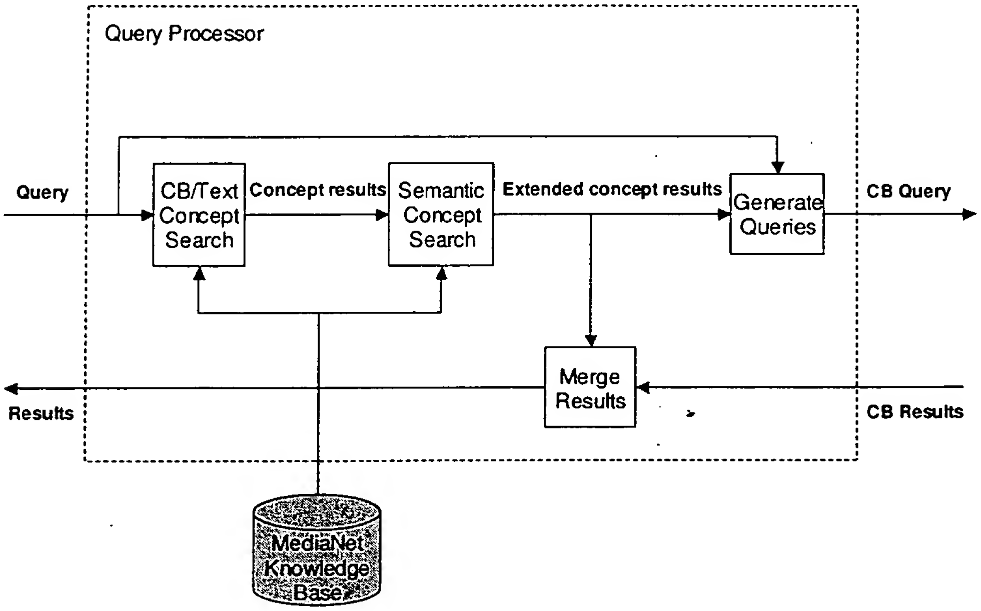


Figure 5: Procedure followed by the Query Processor.

The dissimilarity between two concepts is calculated based on the semantic relationships connecting the two concepts. Dissimilarity scores are assigned to the relationships connecting two concepts as follows. For antonymy, if two concepts are opposite, their dissimilarity is 1; if not, their dissimilarity is zero. For hypernymy/hyponymy, if two concepts are related through the specialization/generalization hierarchy, their dissimilarity is proportional to the number of concepts between the two in the hierarchy; if not, their dissimilarity is 1. For meronymy/holonymy, if two concepts are related through the Has Part,Member,Substance/Is Part,Member,Substance Of network, their dissimilarity is proportional to the number of concepts between the two in the hierarchy; if not, their dissimilarity is 1. The proportional factor for the meronymy/holonymy relationship was set to twice the value of the factor for the hypernymy/hyponymy relationship. In other words, we are giving half the importance to the former relationship compared to the latter one. The total dissimilarity between two concepts is the average dissimilarity for the three relationships. Using this approach, the query can be both expanded and/or refined at the semantic level depending if the new concepts are broader (Generalizes) and/or narrower (Specializes), respectively.

At this point, the query processor has a list of relevant concepts with dissimilarity scores to the query. It sends one query request for the image query and each relevant concept to the CB search engine. The feature vector of the query image is the input of the first query. The feature centroid of the relevant concepts is used as input for the other queries. The query processor merges the results of the multiple queries into a unique list of images and scores by selecting the minimum score for each image among the result lists and shifting the score by the following amount: 0.01 times the dissimilarity score of the concept that generated the minimum score for the image.

For text queries, keywords are match to the list of synonyms of each concept. The feature centroid of the top concept is then treated as an input visual query and the relevant concept. Then, more relevant concepts are found based on conceptual dissimilarity, the queries are submitted to the CB search engine, and the results integrated into a unique list as described above for visual queries.

There is room for improvement in the procedure described above. Multiple relationship paths can connect two concepts could be considered. Then, the conceptual dissimilarity between two concepts could be calculated as the minimum weight of the paths. Single-value decomposition techniques such as latent semantic analysis could be used to obtain distance values between the query, and the database images and/or the concepts in the MediaNet knowledge base. Another new functionality supported a MediaNet knowledge base is browsing and navigation at the concept level. MediaNet also supports advanced inference, problem solving, and decision making tasks that are not implemented in the current retrieval system.

7.3 Experiments

The MediaNet framework has been evaluated in an image retrieval application. The experiments setup and the experimental results are described in the following sub-sections.

7.3.1 Experiments Setup

The objective of these experiments was to evaluate MediaNet in an image retrieval application. We compared the performance of the extended content-based retrieval engine that uses MediaNet to a base retrieval engine that does not use MediaNet. Both retrieval engines used the same image database and CB search engine; however, the MediaNet knowledge base and the query processor were only used in the former (see Figure 4). Two cases were distinguished for the retrieval engine using MediaNet: querying with an image or a keyword. The image retrieval effectiveness [16] was measured in terms of precision and recall. Recall and precision are standard measures used to evaluate the effectiveness of a retrieval engine. Recall is defined as the percentage of relevant images that are retrieved. Precision is defined as the percentage of retrieved images that are relevant.

The image collection selected for the experiment was 5466 images used in MPEG-7 to evaluate and compare color description technology proposed to the standard [25]. This collection includes photographs and frames selected from video sequences from a wide range of domains: sports, news, home photographs, documentaries, and cartoons, among others. The ground truth for 50 queries was also generated by MPEG-7 to compare different color descriptors. For each query, the ground truth represents a semantic, visual class and is annotated by a short textual description (e.g. "Flower Garden" and "News Anchor"). Initially, we used the ground truth generated by MPEG-7 in our experiments to compare the retrieval effectives of both systems but found it to be not suited. We, then, generated the ground truth including relevance scores for the semantic query "tapirs". Relevance scores were assigned to all the images in MPEG-7 ground truth included in the image database as follows: "1" for pictures of tapirs, "0.75" for images of mammals, "0.5" for images of earth animals; "0.25" for images of water and air animals; and "0" for the rest of the images.

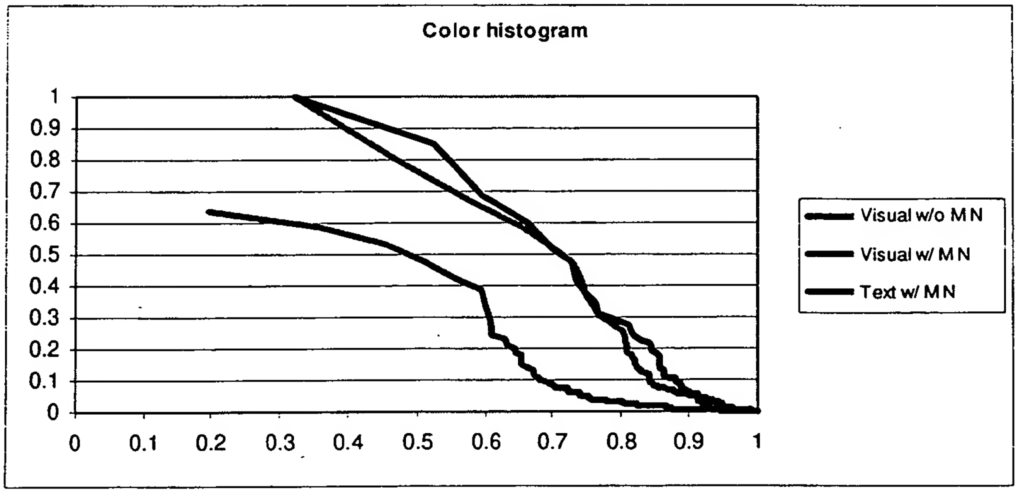
We used the textual descriptions associated with the ground truth of the queries to construct a MediaNet knowledge base as described in section 7.1. The total number of concepts derived from these textual annotations was 96. 50 of these concepts were related to other concepts by generalization/specialization (hypernymy/hyponymy) relationships; 34 concepts were related to other concepts by membership, composition, or substance (meronymy/hyponymy) relationships. There was only one case of antonymy. There were 387 images in the ground truth for all the queries. We generated the image and feature representations of the concepts using half of the images, which were not removed from the image database.

Different experiments were performed with visual queries and text queries as input queries using different feature descriptors. The values P, Q, and R were set to 20%; and the values M and N to 3 (see section 7.1). For visual queries, we used the images specified by MPEG-7 as query inputs. We selected one word from the caption of each class as the input to the text queries. Recall and precision results of the experiments are reported in the next section.

7.3.2 Experimental results

Figure 6 shows the precision and recall for the content-based retrieval and the extended content-based retrieval systems for the 50 queries and the ground truth generated by MPEG-7. For the extended content-based retrieval system, results for text and image queries are provided. In these experiments, both retrieval systems used the color histogram or all for feature descriptors (with equal weights) extracted from the images in the database for retrieval. All the features were normalized to a maximum distance of 1 in the database.

As expected, the retrieval effectiveness for text queries in the extended content-based retrieval engine is considerably lower than for image queries. The results show very similar retrieval effectiveness for image queries by both retrieval systems using color histogram or all the feature descriptors. If two images are retrieved, the recall and the precision values for both systems are about 0.5 and 0.8, respectively. These values are very high and are due to the fact that the ground truth contains images that are very similar visually (contiguous frames from the same video sequence, in most cases) and can be easily and effectively retrieved using low-level features, specially color histogram.



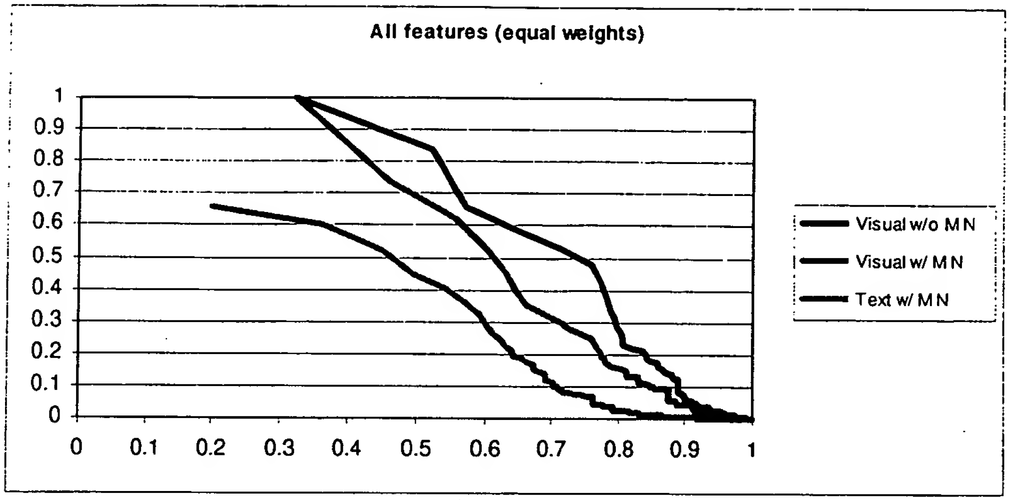
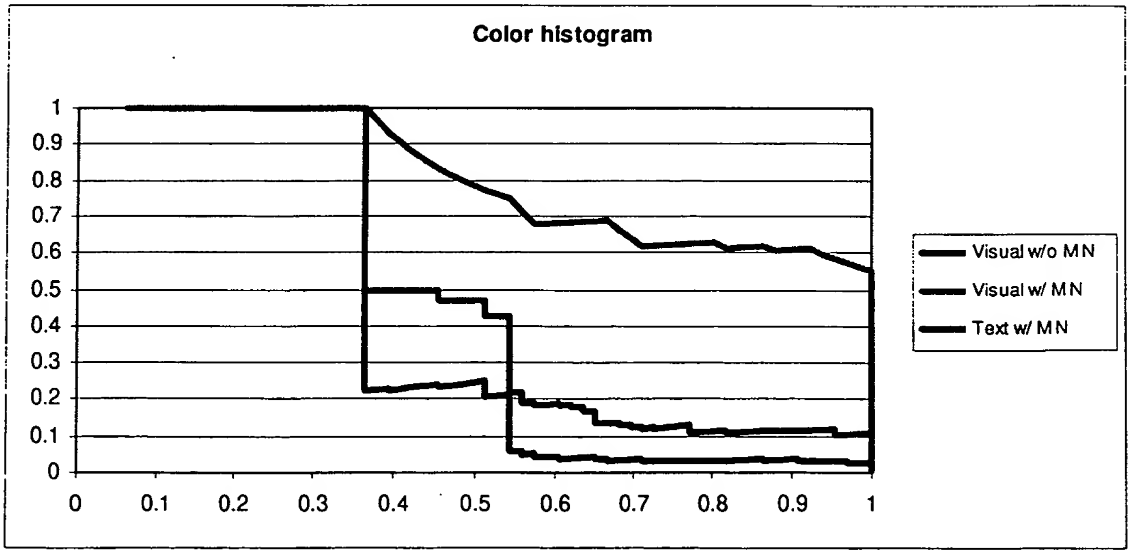


Figure 6: Precision and recall across 50 queries with MPEG-7 ground truth using color histogram (top) and all the features (bottom). "Visual w/o MN" corresponds to the content-based retrieval system that does not use MediaNet; "Visual w/ MN" to the extended content-based retrieval system with image queries; "Text w/ MN" to the extended content-based retrieval system using the text keywords as queries.

Figure 6 made us question the suitability of the ground truth provided by MPEG-7 to evaluate the performance of the extended content-based retrieval. For this reason, we selected the query labeled as "Tapirs" and generated a more semantically meaningful ground truth with scores as described in the previous section. The recall/precision results for the "Tapirs" query are shown in Figure 7. The extended content-based retrieval system for image queries provides consistent precision values over 0.5 for almost any value of recall. The gain of retrieval effectiveness is considerable for both color histogram and all the feature descriptors so we can conclude that the performance gain is due to the MediaNet framework and not the selection of feature descriptors in the retrieval. The results for color histogram are better than for all the color and texture descriptors; the reason for this may be the visual characteristics of the images in the database. These results are quite encouraging but additional experiments are needed to further evaluate the performance gain of a CBIR system using MediaNet.



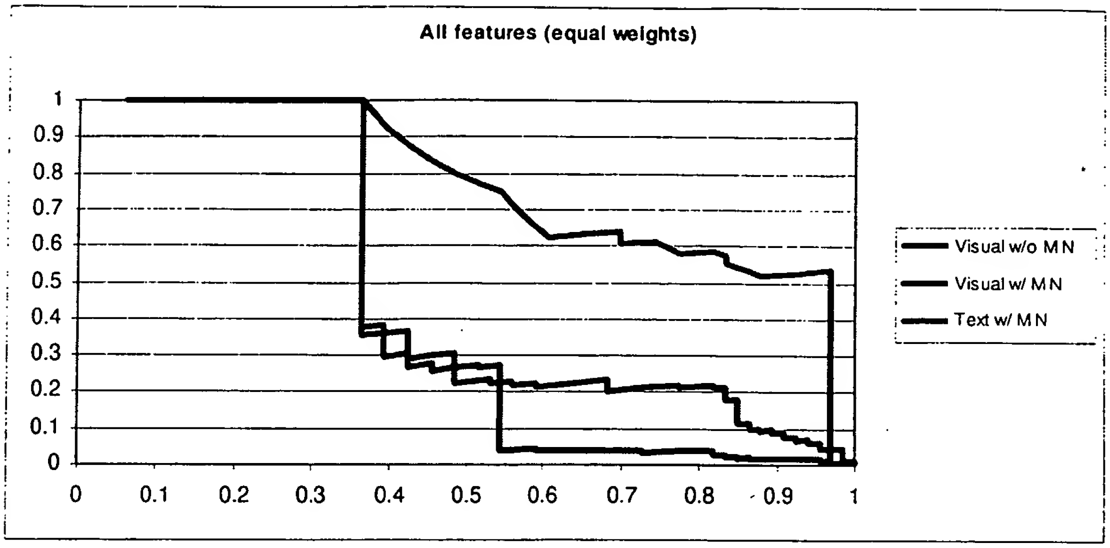


Figure 7: Precision and recall for "Tapirs" query with scored ground truth using color histogram (top) and all the features (bottom). "Visual w/o MN" corresponds to the content-based retrieval system that does not use MediaNet; "Visual w/ MN" to the extended content-based retrieval system with image queries; "Text w/ MN" to the extended content-based retrieval system using the keywords as queries.

8 Conclusions

In this document, we have presented NediaNet, a network of semantic and multimedia information for world knowledge representation, and the encoding, personalization, construction, and usage of MediaNet in applications. We have also described the implementation of the MediaNet framework in a content-based image retrieval system and report on experiments that have demonstrated improved retrieval effectiveness in searching for images using MediaNet.

References

[1] J. R. Bach et al., "Virage Image Search Engine: An Open Framework for Image Management", *Proceeding of Conference on Storage and Retrieval for Image and Video Databases IV (IS&T/SPIE-1996)*, San Jose, California, 1996

[2] Y. A. Aslandogan, C. Their, C. T. Yu, and N. Rishe, "Using Semantic Contents and WordNet in Image Retrieval", *Proc. of the 20th It.l ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 286-295, 1997.

[3] Dewey Decimal System, http://www.oclc.org/dewey/.

[4] M. Dobie, R. Tansley, D. Joyce, M. Weal, P. Lewis, and W. Hall, "A Flexible Architecture for Content and Concept Based Multimedia Information Exploration", *Proc. Of the Challenge of Image Retrieval*, pp. 1-12, Newcastle, Feb. 1999.

[5] M. Flickner et al., "Query by Image and Video Content: The QBIC System", *Computer*, Vol. 28, No. 9, pp. 23-32, Sep. 19995; also available at http://wwwqbic.almaden.ibm.com/.

[6] D. W. Joyce, P. H. Lewis, R. H. Tansley, M. R. Dobie, and W. Hall, "Semiotics and Agents for Integrating and Navigating Through Media Representations of Concepts", *Proc. of Conference on Storage and Retrieval for Media Databases 2000, (IS&T/SPIE-2000)*, Vol. 3972, pp.120-31, San Jose, CA, Jan. 2000.

[7] D. Lenat, "The Dimensions of Context Space", http://www.cyc.com/context-space.doc, Oct. 1998.

[8] W. Y. Ma and B. S. Manjunath, "A Texture Thesaurus for Browsing Large Aerial Photographs", *Journal of the American Society for Information Science (JASIS)*, pp. 633-648, vol. 49, No. 7, May 1998.

[9] A. Meystel, *Semiotic Modeling and Situation Analysis: An Introduction*, AdRem, Bala Cynwyd, PA, 1995.

[10] G. A. Miller, "WordNet: A Lexical Database for English", *Communication of the ACM*, Vol. 38, No. 11, pp. 39-41, Nov. 1995.

[11] MPEG Multimedia Description Scheme Group, "MPEG-7 Multimedia Description Schemes XM (v4.0)", ISO/IEC JTC1/SC29/WG11 MPEG00/N3465, Beijing, CN, July 1999.

[12] MPEG Multimedia Description Scheme Group, "MPEG-7 Multimedia Description Schemes WD (v4.0)", ISO/IEC JTC1/SC29/WG11 MPEG00/N3465, Beijing, CN, July 1999.

[13] MPEG Requirements Group, "MPEG-7: Context, Objectives and Technical Roadmap, V.12", ISO/IEC JTC1/SC29/WG11 MPEG99/N2861, Vancouver, July 1999.

[14] A. Natsev, A. Chadha, B. Soetarman, and J. S. Vitter, "CAMEL: Concept Annotated iMagE Libraries", Submitted.

[15] M. R. Quillian, "Semantic Memory", Semantic Information Processing, M. Minsky (ed), MIT Press, Cambridge, MA, 1968.

[16] J. R. Smith, "Quantitative Assessment of Image Retrieval Effectiveness", To appear in *Journal of Information Access*.

[17] J. R. Smith and A. B. Benitez, "Conceptual Modeling of Audio-Visual Content", *Proc. Intl. Conf. On Multimedia and Expo (ICME-2000)*, July 2000.

[18] J. R. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System", *Proceeding of the ACM Conf. Multimedi*a, ACM Press, New York, 1996; also available at http://www.ctr.columbia.edu/VisualSEEk/.

[19] J. R. Smith and S.-F. Chang, "SaFe: A General Framework for Integrated Spatial and Feature Image Search", *IEEE 1997 Workshop on Multimedia Signal Processing*, 1997.

[20] S. W. Smoliar, J. D. Baker, T. Nakayama, and L. Wilcox, "Multimedia Search: An Authoring Perspective", *Proc. of the First International Workshop on Image Databases and Multimedia Search (IAPR-1996)*, pp. 1-8, Amsterdam, The Netherlands, Aug. 1996.

[21] R. Tansley, "The Multimedia Thesaurus: An Aid for Multimedia Information Retrieval and Navigation", Master Thesis, Computer Science, University of Southampton, UK, Dec. 1998.

[22] R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal, "Automatic the Linking of Content and Concept", *Proc. of the conference on ACM Multimedia 2000*, Los Angeles, CA, Oct. 30 – Nov. 4, 2000.

[23] W3C, "Extensible Markup Language (XML)", http://www.w3.org/XML/.

[24] R. Zhao, and W. I. Grosky, "From Features to Semantics: Some Preliminary Results", *Proc. of IEEE International Conference on Multimedia and Expo 2000*, New York, NY, July 30 – Aug. 2, 2000.

[25] D. Zier, J.-R. Ohm, "Common Datasets and Queries in MPEG-7 Color Core Experiments", ISO/IEC JTC1/SC29/WG11 MPEG99/M5060, Melbourne, Australia, Oct. 1999.

Annex A: Detail Procedure Followed in Creation of MediaNet

Input:

Text annotations for 50 groups of images
Total: 185 images

Output:

The MediaNet knowledge base

Tools:

WordNet, an electronic lexical system for English available on the web (http://www.cogsci.princeton.edu/~wn/obtain/)
Automatic visual feature extraction tools
Human assistance
Perl scripts

Procedure:

1) Generate a list of words by serializing the text annotations.
2) Human assistant removes dummy words such as articles and prepositions ("the" and "of").

Human assistant specifies the syntactic category of each word in the list based on the context of the text annotation.

3) For each word and corresponding syntactic category, use WordNet to generate the senses. The sense description of a word provided by WordNet includes the set of synonyms (synsets) and the definition of the word.
4) For each word, human assistant filters the senses that do not apply in the context of the text annotation and the images annotated with that word.
5) For each pair word and sense, i.e., synset, create a concept in the MediaNet. Assign the sense description as text representations (synonyms and definition) of the content in the MediaNet.

6) For each synset (concept), WordNet was used to generate semantic relationships to others in the MediaNet. The semantic relationships used from WordNet are hypernymy/hyponymy (To Be Specialization/To Be Generalization), meronymy/holonymy (To Have Part, Member, Substance Of/To Be Part, Member, Substance), and antonymy (To Be Opposite).

7) Assign the images associated to each concept as image representations of the concept in the MediaNet.
8) Extract the visual features color histogram, color coherence, tamura texture, and wavelet texture from each image. Assign the extracted features as feature representations of the images in the MediaNet.
9) Assign the centroids of the visual features for all the images associated with a concept as feature representations of the concept in the MediaNet.

Annex B: Detail Procedure Followed in Usage of MediaNet

Input:

Database of color histogram, color coherence, Tamura texture, and wavelet texture features for 5281 images
MediaNet knowledge base for 185 images and text annotations constructed as described above
Image or keyword query

Output:

Ranked list of images in the database visually and semantically similar to the query

Tools:

Automatic visual feature extraction tools
Perl scripts

Procedure for image query:

1) Extract or retrieve the visual feature vector of the query image.
2) Select 3 most relevant concepts in the MediaNet to the query image based on the visual feature vector of the query image and the feature centroid representations of the concepts
    a) Obtain the square Euclidean distance between the visual feature vector of query image and the visual feature centroid of each concept
    b) Select the 20% concepts with minimum distance to the query image
    c) For each selected concept, obtain the square Euclidean distance between the visual feature vector of the query image and the visual feature vector for each image representation of the concept
    d) Select the 20% image representations with minimum distance to the query image
    e) For each selected concept, obtain the distance score to the image query as

$$fdist(q,c) = dist(fv_q, fv\_cen_c) + \sum_{i \in rel_c} \frac{dist(fv_q, fv_i)}{num\_rel_c} * \sqrt{\frac{num_c}{num\_rel_c}}$$

where q is the query image, c is a concept, fv_q is the feature vector of the image query, fv_cen_c is the feature centroid of concept c, num_c is the number of images of concept c, num_rel_c is the number of selected image representations for concept c, rel_c is the set of selected image representations of concept c, i is one image representation, fv_i is the feature vector of the image representation i, and dist(fv_q, fv_cen_c) is the square Euclidean distance among feature vectors fv_q and fv_cen_c.

3) Select at most 3 other concepts in the MediaNet knowledge based on the semantic relationships to the 3 previously selected concepts
    a) Generate a dissimilarity score matrix for the concepts in the MediaNet knowledge based on semantic relationships among them

```
diss_matrix[i,j] = dissimilarity score between concept i and concept j =
    ( hypehypo_diss_matrix[i,j] * hypehypo_weight +
     meroholo_diss_matrix[i,j] * meroholo_weight +
     ant_diss_matrix[i,j] * ant_weight) /
     MAX *( hypehypo_weight + meroholo_weight + ant_weight)
```

where hypehypo_weight = 0.25, meroholo_weight = 0.5, ant_weight = 0.25, and

```
i) hypernymy/hyponymy relationship
      if (concept i and concept j are connected by the hypernymy/hyponymy tree) {
            hypehypo_diss_matrix[i,j] = number of concepts between concepts i and j in the
            hypernymy/hyponymy tree
      } else {
            hypehypo_diss_matrix[i,j] = MAX
      }
ii) meronymy/holonymy relationship
      if (concept i and concept j are connected by the meronymy/holonymy tree) {
```

```
                meroholo_diss_matrix[i,j] = number of concepts between concepts i and j in the
                meronymy/holonymy tree
            } else {
                meroholo_diss_matrix[i,j] = MAX
            }
    iii) antonymy relationship
            if (concept i and concept j are opposite) {
                ant_diss_matrix[i,j] = MAX
            } else {
                ant_diss_matrix[i,j] = 0
            }
```

b) Calculate the average dissimilarity score between each concept in the MediaNet and the three relevant concepts

c) Select at most three concepts with minimum average dissimilarity score greater than the average concept dissimilarity score for all the concepts (0.74)

d) Generate distance scores for the newly selected concepts by summing the average dissimilarity score to the 3 previously selected concepts, and the average distance score between the 3 previously selected concepts and the query image

4) Query the feature database with the feature vector of query image and the feature centroid of the selected concepts and combine the results
   i) Match the feature vector of the query image to the feature database and obtain a ranked list of images in the database. Euclidean distance was used.
   ii) Match the feature centroid of each selected concept to the feature database and obtain a ranked list of images in the database. Euclidean distance was used.
   iii) Generate a unique ranked list of images by selecting the minimum score for each image among the ranked lists resulting from the multiple queries. For the ranked list of images resulting from querying with the feature centroid of the selected concepts, the image scores were shifted by the following amount:

   ```
   0.01 * (dissimilarity score of concept to query image)
   ```

Procedure for text query:

1) Select at most relevant concept in the MediaNet to the query keyword based on text representations of the concepts in the MediaNet
   a) The concept that included the query keyword in their text representation was selected.
2) Select at most 3 other concepts in the MediaNet knowledge based on the semantic relationships to the previously selected concept
   a) Generate a dissimilarity score matrix for the concepts in the MediaNet knowledge based on the semantic relationships among them

```
        diss_matrix[i,j] = dissimilarity score between concept i and concept j =
            ( hypehypo_diss_matrix[i,j] * hypehypo_weight +
             meroholo_diss_matrix[i,j] * meroholo_weight +
             ant_diss_matrix[i,j] * ant_weight) /
             MAX *( hypehypo_weight + meroholo_weight + ant_weight)
    where hypehypo_weight = 0.25, meroholo_weight = 0.5, ant_weight = 0.25, and
    i)  hypernymy/hyponymy relationship
            if (concept i and concept j are connected by the hypernymy/hyponymy tree) {
                hypehypo_diss_matrix[i,j] = number of concepts between concepts i and j in the
                hypernymy/hyponymy tree
            } else {
                hypehypo_diss_matrix[i,j] = MAX
            }
    ii) meronymy/holonymy relationship
            if (concept i and concept j are connected by the meronymy/holonymy tree) {
                meroholo_diss_matrix[i,j] = number of concepts between concepts i and j in the
                meronymy/holonymy tree
            } else {
```

```
                meroholo_diss_matrix[i,j] = MAX
            }
    iii) antonymy relationship
            if (concept i and concept j are opposite) {
                ant_diss_matrix[i,j] = MAX
            } else {
                ant_diss_matrix[i,j] = 0
            }
```

b) Calculate the average dissimilarity score between each concept in the MediaNet and the selected concept

c) Select at most three concepts with minimum average dissimilarity score greater than the average concept dissimilarity score for all the concepts (0.74)

d) Generate distance scores for the newly selected concepts as the average dissimilarity scores of each newly selected concept to the previously selected concept

3) Query the feature database with the feature centroid of the selected concepts and combine the results

i) Match the feature centroid of each selected concept to the feature database and obtain a ranked list of images in the database. Euclidean distance was used.

ii) Generate a unique ranked list of images by selecting the minimum score for each image in the ranked lists resulting from the multiple queries. For the ranked list of images resulting from querying with the feature centroid of the selected concepts, the image scores were shifted by the following amount:

0.01 * (dissimilarity score of concept to query image)